ON THE VARIANCE OF THE OPTIMAL ALIGNMENT SCORE FOR AN ASYMMETRIC SCORING FUNCTION

Christian Houdré * and Heinrich Matzinger†

Abstract

We investigate the variance of the optimal alignment score of two independent iid binary, with parameter 1/2, sequences of length n. The scoring function is such that one letter has a somewhat larger score than the other letter. In this setting, we prove that the variance is of order n, and this confirms Waterman's conjecture in this case.

1 Introduction

The problem under investigation in this article is similar to the corresponding problem for the Longest Common Subsequence (LCS), which is, in fact, a special case of optimal alignment. It is well known that the determination of the order of the fluctuations for the LCS-problem is one of the main problem in computational biology, and to this day it remains open. As it will be explained below, the optimal alignment problem can be formally viewed as a last passage percolation problem with correlated weights. For first and last passage percolation the exact order, in the general case, is still unknown. In some special cases it has been found as well as for the Longest Increasing Subsequence (LIS) problem, where the exact asymptotic distribution is known. In turn, the LIS problem is itself asymptotically equivalent to a first passage percolation problem on a Poisson graph.

We will return later to the history of these questions and of related ones but, first, let us explain the applications we have in mind. Optimal alignment has gained tremendous significance in both computational biology and computational linguistics. The reader will find in the standard references [9], [11], [19], and [20] for a general discussion of the relevance of string comparison, and related problems, in computational biology.

*School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.
Email: houdre@math.gatech.edu

†University of Bielefeld Postfach 10 01 31, 33501 Bielefeld, Germany.
Email: matzing@mathematik.uni-bielefeld.de

Optimal alignments and closely related methods are one of the main tools for identifying genes. For example, with the help of optimal alignment one searches for the location of gene X. Assume that we have already identified the location of related genes. The related genes will exhibit a certain amount of similarity. We hunt through the genome in the hope of finding a spot that resembles X. To evaluate the degree of similarity, we use an alignment score. A high alignment score indicates high similarity.

In computational linguistics, a quite common task consists in determining, typically across languages, related pieces of texts or similar words. Often, one needs to identify pairs of translated words as this is an important step towards building electronic lexicons or a translation machine. Suppose we are given two texts: the first text is in French, while the second text is in German. The two texts are translations of one another and we try to build a computer program able to determine which words are translation of one another. Many translation pairs exhibit great similarities. Our program should be able to automatically detect such similarities without prior knowledge of the languages. Let us present a concrete English/German example: Let X be the English word *brother* and let Y be the German translation $Y = bruder$. Looking at the two words, we immediately observe a great degree of similarity between them. The computer should also be able to detect this resemblance. A first, unsophisticated method consists in writing one word beneath the other and to count the number of coinciding letters. We then find:

b	r	o	t	h	e	r
b	r	u	d	e	r	

Two letters coincide: both words start with the letters *br*. This is not yet very conclusive. For example the words brag, bread, breast, bribe, bride, bring, broad, brute, all start with the letters *br*. Hence this method of alignment is not very powerful to help discriminate unrelated pairs of words across similar languages. A better solution consists in aligning the two words allowing for gaps whilst trying to obtain the maximum possible number of coinciding letters. With this method, the *optimal alignment* turns out to be:

b	r			o	t	h	e	r
b	r	u	d				e	r

This time we get a sequence of four coinciding letters: *brer*. Note that *brer* is a common subsequence of X and Y. This means that the word *brer* can be obtained from both X as well as from Y by only deleting letters. It turns out that *brer* is the Longest Common Subsequence (LCS) of X and Y. (This means that *brer* is the longest sequence which is a subsequence of both X and Y). A further improvement consists in also allowing the alignment of similar letters. For example, give a score of one for identical letter, but a score of 0.5 when the letters are only similar. Assuming that t and d are similar, and that so are o and u, we find the following optimal alignment:

b	r	o	t	h	e	r
b	r	u	d		e	r

(1.1)

The score of the above alignment is $4 + 2(0.5) = 5$, since four identical letters are aligned as well as two similar ones. The score of the alignment (1.1) can be written as:

$$s(b,b) + s(r,r) + s(o,u) + s(t,d) + s(e,e) + s(r,r) = 4 + 2(1/2) = 5,$$

where $s(x,y)$ denotes the score obtained by aligning the letter x with the letter y. Sometimes a gap penalty is also in use. Note that the length of the LCS is equal to the optimal alignment score when the substitution matrix s is taken to be the identity matrix and that a zero gap penalty is in force. Summarizing: the total score of an alignment is the sum of terms for each aligned pair of residues, plus a usually negative term for each gap, the so-called gap penalty.

Throughout this paper the texts X and Y are iid binary random text independent of each other and both have length n. The optimal alignment score is denoted by L_n. The substitution matrix is such that $s(1,1)$ is somewhat larger than the other single-letter scores. In practice this corresponds to a situation where one letter is more stable than the other letters. The main result of the present paper is that $\mathrm{Var}L_n$ is of order n, when the score $s(1,1)$ is taken large enough.

As already mentioned, the study of the fluctuations for the LCS is one of the main problems in computational biology. For the LCS-case, Steele [21], [22] proved that $\mathrm{Var}L_n \leq n$, while Chvàtal and Sankoff conjectured that $\mathrm{Var}L_n$ is of order $o(n^{2/3})$, which is the same order (when properly rescaled), as the one obtained by Baik, Deift and Johansson [8] in their much celebrated result on the Longest Increasing Subsequence (LIS) of a random permutation. (See, [**?**], for more precise references to this problem). In turn, Waterman [23] conjectured that in many cases the variance of L_n grows linearly. A similar order is obtained by [16] when one sequence is not random but rather periodic.

In the LIS-problem, one picks randomly a permutation of $\{1, 2, \ldots, n\}$ and considers the LIS of the sequence of permutated numbers. The LIS-problem is asymptotically equivalent to a Last Passage Percolation (LPP) problem on an oriented Poisson graph. This corresponds to collecting as many Poisson points, in a square, when one is only allowed to move rightward or upward. The Poisson point process is then homogeneous. As we will argue later, the optimal alignment problem is equivalent to a Last Passage Percolation (LPP) problem with correlated weights. For First Passage Percolation (FPP) and LPP, the order of the (longitudinal) fluctuation is not yet know in full generality, but some interesting results are available. Our result, also proves that the behavior of optimal alignment scores can be entirely different from the typical LPP and LIS-situations, despite the fact that optimal alignment is formally a LPP-problem.

Let us quickly explain how optimal alignment can be viewed as a LPP-problem. The LPP problem consists in going in an oriented graph from point A to point B. Each edge e is assigned a weight $w(e)$, which represents the time it takes to cross e. In the LPP problem, one is interested in finding the path which takes the most time to go from A to B. (In First Passage Percolation (FPP) one searches for the shortest path on a non-oriented graph). In the optimal alignment problem, the set of vertices V is $\mathbb{N}^2$. Let $\vec{e}_1 := (1,0), \vec{e}_2 := (1,1), \vec{e}_3 := (0,1)$ and let $E_A := \{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$. Let the weight of the edge $w(\vec{x}, \vec{y})$ be equal to the gap penalty when $\vec{y} - \vec{x} = \vec{e}_1$ or $\vec{y} - \vec{x} = \vec{e}_3$. When, $\vec{y} - \vec{x} = \vec{e}_2$, then $w(\vec{x}, \vec{y})$ is equal to the alignment score $s(X_i, Y_j)$ where $\vec{y} = (i,j)$ and the aligned texts are $X = X_1 X_2 \ldots X_n$ and $Y = Y_1 Y_2 \ldots Y_n$. Note that here the weights are correlated which is in stark contrast to classical LPP and FPP where the weights are mostly taken to be iid. In the optimal alignment problem the weights $s(X_i, Y_j)$ are function of two one dimensional texts. Hence, unlike FPP and LPP the optimal alignment problem is an hybrid problem with some aspects being two-dimensional and others being one-dimensional.

Let us now discuss a little bit more some of the history of LCS-problems. Using a subadditivity argument, Chvàtal and Sankoff [10] prove that the limit

$$\gamma := \lim_{n\to\infty} \frac{E[L_n]}{n}$$

exists, where L_n is the length of the LCS of two independent iid sequences of length n. To this day, the value of the Chvàtal and Sankoff constant $\gamma > 0$, is unknown, even for binary sequences. Waterman and Arratia [6] derive a law of large deviations for the fluctuations of L_n on scales larger than $\sqrt{n}$. In this groundbreaking article, they show the existence of a critical phenomenon. Using first passage percolation methods, Alexander [2] proves that $E[L_n]/n$ converges to γ at a rate of order at least $\sqrt{\log n/n}$. In [23], Waterman studies the statistical significance of the results produced by sequence alignment methods.

The determination of the order of the the variance for the LCS problem is what will be of concern to us in the rest of the text.

2 The Main Result

Throughout this paper $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ are two independent iid sequences of Bernoulli random variables with parameter 1/2. We also assume in everything that follows that the substitution matrix is such that:

$$s(1,0) = s(0,1) = 0 \text{ and } s(0,0) = 1, \tag{2.1}$$

while the gap penalty is taken equal to zero:

$$q = 0 \tag{2.2}$$

We consider the two sequences of equal length $X := X_1X_2\dots X_n$ and $Y := Y_1Y_2\dots Y_n$. An *alignment* is a pair of increasing sequences (π,ν) such that

$$1 \le \pi(1) < \pi(2) < \dots < \pi(k) \le n$$

and

$$1 \le \nu(1) < \nu(2) < \dots < \nu(k) \le n,$$

where $\pi = \pi(1)\pi(2)\dots\pi(k)$, $\nu = \nu(1)\nu(2)\dots\nu(k)$ and $k \le n$. The score of the alignment (π,ν) is defined as

$$S_{(\pi,\nu)} := \sum_{i=1}^{k} s(X(\pi(i)), Y(\nu(i))).$$

The *optimal score* L_n is defined by

$$L_n := \max S_{(\pi,\nu)},$$

where the maximum is taken over all possible alignments (π,ν).

The main result of the present paper asserts that if the score $s(1,1)$ is large enough, then the variance of the optimal score is of order n. More precisely,

Theorem 2.1 *Let $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ be two independent sequences of iid Let $0 \le i < j \le n$. The interval $[i,j]$ is called a block of zeros in X if*

$$X_i = X_{i+1} = \ldots = X_j = 0,$$

but $X_{i-1} = 1$ (or $i = 1$) and $X_{j+1} = 1$ (or $j = n$). We integer $j-i+1$ is then called length of the block $[i,j]$.Bernoulli random variables with parameter $1/2$. Let the substitution matrix be such that $s(1,0) = s(0,1) = 0, s(0,0) = 1$ with moreover no gap penalty, i.e., $q = 0$. Then, there exist $c > 0$ and $s > 0$, such that if $s(1,1) \ge s$, then

$$\mathrm{Var} L_n \ge c \cdot n. \tag{2.3}$$

for all $n \in \mathbb{N}$.

The last theorem implies that $\mathrm{Var}L_n = \Theta(n)$, using the upper bound obtained, via the tensorization property of the variance, in [21], [22].

The main idea developed in proving the above theorem, is to show that when changing the length of a randomly chosen block in X, then this process has a tendency to increase the score. Since the number of blocks of a certain length has fluctuations of order $\Theta(n)$ this, in turn, implies that $\mathrm{Var}L_n$ is of order $\Theta(n)$.

Let us now explain what a block of X is, starting with a formal definition.

Let $0 \le i < j \le n$. The interval $[i,j]$ is called a *block of zeros* (resp. *block of ones*) in X if

$$X_i = X_{i+1} = \ldots = X_j = 0, (resp. = 1)$$

but $X_{i-1} = 1 (resp. 0)$ (or $i = 1$) and $X_{j+1} = 1 (resp. 0)$ (or $j = n$). The integer $j - i + 1$ is then called the *length of the block* $[i,j]$.

Here is a numerical example: Let X be equal to the string: 000011100000. The first block in this case consists of four zeros and has length four. The second block consists of three ones and has length three. The third block is a block of zeros of length five

Next, let us described our transformation: We pick a block of zeros of length five at random among all the blocks of zeros of length five in X. For this we use the equiprobable distribution and this selection process is independent of Y. Then, we take off one zero from the chosen block. The block becomes a block of length four. The next step is to add the zero just removed to a randomly chosen block of zeros of length one. This block then becomes a block of length two. Again, to chose the block of length one, we use the equiprobable distribution on all blocks of zeros and length one in X. The string X gets transformed in this way into a new string with same length. This new string is denoted by $\tilde{X}$.

Let us, once more, consider a numerical example. Let $X = 0101000001000001$. The string X has two blocks of zeros of length one as well as two blocks of zeros of length five. Assuming that the second block of length five gets chosen and so is the first block of length one, we would obtain: $\tilde{X} = 001010000010000$.

The optimal alignment score of $\tilde{X}$ with Y is denoted by $\tilde{L}_n$. Hence:

$$\tilde{L}_n := \max \sum_{i=1}^{k} s(\tilde{X}(\pi(i)), Y(\nu(i))),$$

where the maximum is taken over all alignments (π, ν).
We show that, when $s(1,1)$ is taken large enough, then $\tilde{L}_n$ tends to be larger than L_n. This is the content of the next theorem:

Theorem 2.2 *Assume that* (2.1) *and* (2.2) *both hold. Then, there exist* $\epsilon_1, c_1, s > 0$ *and* $s > 0$ *all independent of* n*, such that if* $s(1,1) \geq s$*, then*

$$P(A^n) \geq 1 - e^{-c_1 n}, \tag{2.4}$$

for all $n \in \mathbb{N}$*, where* A^n *is the event that the following two equations both hold:*

$$P(\tilde{L}_n - L_n = 1 | X, Y) \geq \frac{31}{32} \cdot \frac{1}{4} - \epsilon_1, \tag{2.5}$$

$$P(\tilde{L}_n - L_n = -1 | X, Y) \leq \frac{1}{32} + \epsilon_1. \tag{2.6}$$

Let us explain the main idea behind the previous theorem: when $s(1,1)$ is large most of the ones will get matched with ones. So, we are close to a situation where we match all the ones and match as many zeros in between ones as possible. For an alignment which tries to match all the ones, the blocks of zeros between matched ones are i.i.d.. The distribution of the length of the blocks of zeros between matched ones is geometric with parameter 1/2.

Let us look at another numerical example of an alignment obtained by matching all the ones. Take $X = 1011000001$ and $Y = 10010101$. When we match all the ones and as many zeros as possible in between matched ones, we obtain the following alignment:

1	0	_	1	_	1	0	0	0	0	0	1
1	0	0	1	0	1	0	_	_	_	_	1

In our current example, the first block of zeros of X has length 1 and the third has length 5. If we shorten the third block by removing a zero and adding it to the first block the alignment-score gets increased by one. The reason for this is that the third block of zeros of X is matched with a shorter block so removing a zero does not reduce the score. However the first block of zeros of X is matched with a longer block of Y so adding a zero increases the score by one unit. When we take a bit from the third block to add it to the first block of zeros of X, we get the following new alignment:

1	0	0	1	_	1	0	0	0	0	1
1	0	0	1	0	1	0	_	_	_	1

This alignment has score 7. This is one unit larger than the score of the alignment before modifications.

In general the alignment which aligns all ones is not an optimal alignment. It is very difficult to understand how the optimal alignment looks macroscopically for long texts. There are extremely complicated dependencies between all the parts of the optimal alignment. But despite the horrendously complicated macroscopic behavior, when most ones are matched, the local distribution of the optimal alignment is close to the alignment where we match all the ones. Furthermore, the alignment where we match all the ones consists of a sequence of iid blocks of zeros between matched ones.

For the alignment with all ones matched, we have that: We can compute the probability that the score increases by one when we take a zero from a block of X of length five and add it to a block of length one. The block are chosen at random. The corresponding blocks in Y are iid and have geometric distribution. Hence the probability to have the alignment score increased by one is the probability that the chosen block of length five is matched with a block of shorter length times the probability that the chosen block of length one is matched with a longer block. Hence the probability of an increase in score is:

$$P(Z < 5) \times P(Z > 1) = \frac{31}{32} \cdot \frac{1}{4},$$

where Z is a geometric random variable with parameter $1/2$.

Similarly, the probability that the score decreases by one unit is the probability that the block of length five is matched to a block of length five or longer, times the probability that the block of length one is matched to a block of length one or no zero. Hence the probability that the score decreases by one unit is

$$P(Z \geq 5) \times P(Z \leq 1) = \frac{3}{32 \cdot 4}.$$

The probability to have the score unchanged is:

$$P(Z \geq 5) \times P(Z > 1) = \frac{1}{32 \cdot 4} - \frac{3}{4}\frac{31}{32}.$$

Theorem 2.2 says that the probabilities for the optimal score to increase/decrease through our bit transfer procedure, is close to the probabilities for the alignment which matches all the ones. (For this we assume that most ones are matched due to the score $s(1,1)$ being large). Theorem 2.2 is proved in Section 4 and Section 5 using exponential estimates and some combinatorics. In the next section, we prove that Theorem 2.2 implies Theorem 2.1.

3 Proof of the Main Theorem

The purpose of section is to prove that Theorem 2.2 implies our main Theorem 2.1. We first need a few definitions.

Let N_i be the number of blocks of zeros of length i in the string $X = X_1 X_2 \ldots X_n$, and let $\vec{N}$ be the vector $\vec{N} = (N_1, N_2, N_4, N_5)$. Let $H_a = \{(n_1, 0, n_4, n_5) | n_1, n_4, n_5 \in \mathbb{N}\}$ and let $H_b = \{(n_1, n_2, 0, n_5) | n_1, n_2, n_5 \in \mathbb{N}\}$. Let $\vec{e} := (-1, 1, 1, -1)$. It is easy to check that every integer vector $\vec{n} = (n_1, n_2, n_4, n_5) \in \mathbb{N}^4$ can be written in a unique way as a vector of $H_a \cup H_b$ plus a term $n \cdot \vec{e}$ where $n \in \mathbb{N}$.

For every $\vec{n} \in H_a \cup H_b$, we define a sequence of random strings:

$$X(\vec{n}), X(\vec{n} + \vec{e}), X(\vec{n} + 2\vec{e}), \ldots, X(\vec{n} + k\vec{e}), \ldots$$

The sequence is defined by induction on k:

-First, $X(\vec{n})$ is a binary random string of length n having distribution $\mathcal{L}(X | \vec{N} = \vec{n})$. We also ask that $X(\vec{n})$ be independent of Y.

-Once, $X(\vec{n}+k\vec{e})$ is defined, let $X(\vec{n}+(k+1)\vec{e})$ be the string obtained by taking one zero from a block of length five and adding it to a block of length 1. For this we draw one block among all the blocks of zeros of $X(\vec{n}+k\vec{e})$ of length five and then we pick another block of zeros of length one in $X(\vec{n}+k\vec{e})$ and turn it into a block of length two. The blocks are drawn with equal probability among all blocks of length five, resp. one.

Let us look, once again, at a numerical example. Take $n = 16$ and $\vec{n} = (2, 0, 0, 2)$. Hence the string $X(\vec{n})$ in this case, is a binary random string of length 16. It distribution is the conditional distribution of an iid string, given that there are:
-two blocks of zeros of length one,
-no blocks of zeros of length two or four,
-two blocks of zeros of length five.
We generate the string $X((2, 0, 0, 2))$ with this distribution. We could get for example

$$X((2, 0, 0, 2)) = 0101000001100000.$$

To obtain the next string, that is $X((1, 1, 1, 1))$ we pick one of the blocks of length five at random and turn it into a block of length four. Then, we add a zero to one of the block of length one. There are two blocks of length five, so they have both equal probability to get chosen. Assume that the randomly chosen block of length five is the second one. Assume that the randomly chose block of length one turns out to be the first block. Then, we would have that

$$X((1, 1, 1, 1)) = 0010100000110000.$$

Let $L(\vec{n})$ be the optimal alignment score of $X(\vec{n})$ and Y.

The next lemma says that for any $\vec{m} \in \mathbb{N}^4$, such that

$$P(\vec{m} = \vec{N}) > 0,$$

we have that $L(\vec{m})$ has the distribution of L_n conditional on $\vec{N} = \vec{m}$.

Lemma 3.1 *Let $\vec{m} \in \mathbb{N}^4$ be such that*

$$P(\vec{m} = \vec{N}) > 0.$$

Then, $L(\vec{m})$ has distribution

$$\mathcal{L}(L_n|\vec{m} = \vec{N}).$$

Proof. We need to prove that for any $\vec{m} \in \mathbb{N}^4$ with $P(\vec{N} = \vec{m}) > 0$, we have that

$$\mathcal{L}(X(\vec{m})) = \mathcal{L}(X|\vec{N} = \vec{m}).$$

The proof is by induction on k. By definition, we have that $X(\vec{n})$ has distribution $\mathcal{L}(X|\vec{N} = \vec{n})$ when $\vec{n} \in H_a \cup H_b$.
The next step is to prove that when $X(\vec{n}+k\vec{e})$ has distribution:

$$\mathcal{L}\left(X|\vec{N} = \vec{n}+k\vec{e}\right),$$

then $X(\vec{n}+(k+1)\vec{e})$ must have distribution

$$\mathcal{L}\left(X|\vec{N}=\vec{n}+(k+1)\vec{e}\right).$$

Let $\vec{m}\in\mathbb{N}^4$ be such that $P(\vec{N}=\vec{m})\neq 0$. The distribution $\mathcal{L}(X|\vec{N}=\vec{m})$ can be characterized as the equiprobability distribution on the set of strings x for which

$$\vec{N}(x)=\vec{m}.$$

Here $\vec{N}(x)\in\mathbb{N}^4$ is the vector $(n_1^x,n_2^x,n_4^x,n_5^x)$, where n_i^x designates the number of blocks of zeros of length i in the string x.
Let $\vec{n}^*$ be the vector $\vec{n}^*:=\vec{n}+k\vec{e}$. Let

$$\vec{n}^*=(n_1^*,n_2^*,n_4^*,n_5^*).$$

So we only need to prove that all the possible realizations of $X(\vec{n}^*+\vec{e})$ are equiprobable and that for every x in the set

$$\{x\in\{0,1\}^n|\vec{N}(x)=\vec{n}^*+\vec{e}), \tag{3.1}$$

the probability

$$P(X(\vec{n}^*+\vec{e})=x),$$

is non zero.
Any string x, such that $\vec{N}(x)=\vec{n}^*+\vec{e}$ has a non-zero number of blocks of zeros of length 2 and of length 4. Take any of these blocks of length 4 and any block of length 2. Let y be the string obtained by reducing the chosen block of length two by one bit and adding this bit to the block of length 4. By induction assumption, the probability for $X(\vec{n}^*)$ to be equal to y is non-zero. The conditional probability, given $X(\vec{n}^*)=y$, that $X(\vec{n}^*+\vec{e})=x$ is equal to $1/(n_5^*n_1^*)$ which is different from zero. Hence, $P(X(\vec{n}^*+\vec{e})=x)$ is larger than zero since

$$P(X(\vec{n}^*+\vec{e})=x)\geq\frac{P(X(\vec{n}^*)=y)}{n_5^*n_1^*}.$$

It remains to prove that for all x such that

$$\vec{N}(x)=\vec{n}^*+\vec{e},$$

we have that $P(X(\vec{n}^*+\vec{e})=x)$ does not depend on x. By law of total probability we have that

$$P(X(\vec{n}^*+\vec{e})=x)=\sum_y P(X(\vec{n}^*)=y)\cdot P(X(\vec{n}^*+\vec{e})=x|X(\vec{n}^*)=y). \tag{3.2}$$

The sum in the equality above is taken over all strings y which can be transformed into the string x by taken a bit from a block of zeros of length five and adding it to a block of length one. For each pair consisting of a block of zeros of length four and a block of zeros of length two of x there is such a y. Hence the number of terms in the sum on the right side of equality 3.2, is equal to

$$n_2^x \times n_4^x = (n_2^* - 1)(n_4^* - 1).$$

By the induction assumption, for a string y such that $\vec{N}(y) = \vec{n}^*$, we have that $P(X(\vec{n}^*) = y)$ does not depend on y.
The probability for y to get transformed into x, is equal to:

$$P(X(\vec{n}^* + \vec{e}) = x | X(\vec{n}^*) = y) = \frac{1}{n_1^* n_5^*}.$$

Hence equality 3.2 becomes:

$$P(X(\vec{n}^* + \vec{e}) = x) = P(X(\vec{n}^*) = y) \cdot \frac{(n_2^* + 1)(n_4^* + 1)}{n_1^* n_5^*}, \tag{3.3}$$

where y is any string such that $\vec{N}(y) = \vec{n}^*$. As already mentioned, by the induction assumption, for such a y, we have that $P(X(\vec{n}^*) = y)$ does not depend on y. Hence, the expression on the right side of equation 3.3 does not depend on x. This finishes the proof. ■

We assume that the random variables $L(\vec{n})$ where constructed so that they are independent of $\vec{N}$. Then $L(\vec{N})$ has same distribution as L_n. Hence,

$$\mathrm{Var}[L_n] = \mathrm{Var}[L(\vec{N})] \tag{3.4}$$

We mentioned that any vector of $\mathbb{N}^4$ can be represented in a unique way as a sum: $\vec{n} := m_1 + k\vec{e}$. Where $m_1 \in H_a + H_b$ and $k \in \mathbb{N}$. Let $\vec{M} \in H_a \cup H_b$ and $M_1 \in \mathbb{N}$ be defined by the equation:

$$\vec{N} = \vec{M} + M_1 \vec{e}. \tag{3.5}$$

We also write $L(M_1, \vec{M})$ for $L(\vec{N})$.
Let H^n be the intersection of $H_a \cup H_b$ with $B(0, \sqrt{n})$. (Here $B(0, \sqrt{n})$ stands for the ball of radius $\sqrt{n}$ centered at 0.)
Let $H_e^n := \{k\vec{e} | k \in [-\sqrt{n}, \sqrt{n}]\}$. Let $\vec{\mu} := (\mu_1, \mu_2, \mu_4, \mu_5)$ where

$$\mu_1 := E[N_1], \mu_2 := E[N_2], \mu_4 := E[N_4], \mu_5 := E[N_5].$$

Let

$$I^n := C \cdot H^n + C \cdot H_e^n + \vec{\mu}, \tag{3.6}$$

where $C > 0$ is a constant not depending on n, chosen large enough so that

$$\vec{\mu} + [-\sqrt{55n/4}, \sqrt{55n/4}]^4 \subset I^n, \tag{3.7}$$

holds. Let E_{slope}^n be the event that for all $\vec{n}_a, \vec{n}_b \in I^n$ such that $\vec{n}_b - \vec{n}_a = k\vec{e}$, with $k \geq n^{0.1}$ we have that

$$L(\vec{n}_b) - L(\vec{n}_a) \geq 0.01 |\vec{n}_b - \vec{n}_a|.$$

For any random variables V and W, we write $\mathrm{Var}[V|W]$ for $E[V^2|W]-(E[V|W])^2$. We have that

$$\mathrm{Var}[V]=E[\mathrm{Var}[V|W]]+\mathrm{Var}[E[V|W]],$$

and hence

$$\mathrm{Var}[V]\geq E[\mathrm{Var}[V|W]].$$

This give in our case,

$$\mathrm{Var}[L(\vec{N})]\geq E\left[\ \mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\ \right]. \tag{3.8}$$

Since the variance is positive, we find

$$E\left[\ \mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\ \right]\geq \tag{3.9}$$

$$\geq E\left[\ \mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\ \Big|\ \vec{N}\in I^n,E^n_{\mathrm{slope}}\right]\cdot P(\vec{N}\in I^n)\cdot P(E^n_{\mathrm{slope}}) \tag{3.10}$$

Note that when we condition on $\vec{M}$ and hold $\vec{M}$ fixed, then $L(M_1,\vec{m})$ becomes a function of the one dimensional variable M_1. When the event E^n_{slope} holds that function has "positive slope on the scale $n^{0.1}$". We are hence interested in the variance of a non-random function of a random variable. For this consider:
Assume that $f(.):\mathbb{R}\rightarrow\mathbb{R}$ is map such that $f'(x)>c$ for all $x\in\mathbb{R}$. (Here $c>0$ is a constant not depending on x). Then, for any random variable Y, we have

$$\mathrm{Var}[f(Y)]\geq c^2\mathrm{Var}[Y]. \tag{3.11}$$

Hence, if the map $x\mapsto L(x,\vec{M})$ would have partial derivative along x everywhere larger than $c>0$, is would follow that $\mathrm{Var}[L(M_1,\vec{m})]$ is larger than $c^2\cdot\mathrm{Var}[M_1|\vec{M}]$. Typically, the integer map $x\mapsto L(x,\vec{M})$ does not strictly increase in the direction of $\vec{e}$. But it is likely that the event E^n_{slope} holds, and hence there is a linear increase in the direction of $\vec{e}$ every $n^{0.1}$ points.
We are next going to formulate a lemma which is a modification of inequality 3.11, for when the map $f(\cdot)$ does not increase every k, but has a tendency to increase on some scale:

Lemma 3.2 *Let $c,m>0$ be two constants. Let $f:\mathbb{Z}\rightarrow\mathbb{Z}$ be a non decreasing map such that:*

- *for all $i<j$:*

$$f(j)-f(i)\leq(j-i) \tag{3.12}$$

- *for all i,j such that $i+m\leq j$:*

$$f(j)-f(i)\geq c\cdot(j-i). \tag{3.13}$$

Let B be an integer random variable such that $E[|f(B)|]\leq+\infty$. Then:

$$\mathrm{Var}[f(B)]\geq c^2\left(1-\frac{2m}{c\sqrt{\mathrm{Var}[B]}}\right)\mathrm{Var}[B]. \tag{3.14}$$

We can now use the result of lemma 3.2 to give a lower bound for $\mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]$ when E^n_{slope} holds and assuming that we condition on $\vec{N}\in I^n$. When E^n_{slope} holds, the integer map

$$x\mapsto L(x,\vec{M})$$

restricted to $\{x|x\vec{e}+\vec{M}\in I^n\}$ satisfies the conditions of lemma 3.2 with $m=n^{0.01}$ and $c=0.01$. Hence, we obtain that conditional on $\vec{N}\in I^n$ and when E^n_{slope} holds, then

$$\mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\geq 0.01^2\left(1-\frac{2n^{0.1}}{0.01\sqrt{\mathrm{Var}[M_1|\vec{M},\vec{N}\in I^n]}}\right)\mathrm{Var}[M_1|\vec{M},\vec{N}\in I^n]. \tag{3.15}$$

Next we need the following lemma:

Lemma 3.3 *Assume that W is a random variable taking its values in $\mathbb{Z}$:*

$$P(W\in\mathbb{Z})=1.$$

Assume that $\kappa>0$ is a constant not depending on n. Let J be an interval of diameter at least $3\kappa\ln 2\sqrt{n}$, such that $P(W\in J)=1$. Assume that

$$P(W=i+1)\geq P(W=i)\cdot\left(1-\frac{\kappa}{\sqrt{n}}\right) \tag{3.16}$$

and

$$P(W=i)\geq P(W=i+1)\cdot\left(1-\frac{\kappa}{\sqrt{n}}\right) \tag{3.17}$$

for all $i,i+1\in J$.
Then,

$$\mathrm{Var}[W]\geq n\cdot\frac{(\ln 2)^2}{16\kappa^2}. \tag{3.18}$$

for every n large enough.

Proof. Let I be the interval

$$I:=[\ E[W]-\sqrt{n}\ln 2/(2\kappa)\ ,\ E[W]+\sqrt{n}\ln 2/(2\kappa)\ ].$$

Let I_r, resp. I_l denote the interval of length $\ln 2\sqrt{n}/\kappa$ directly adjacent to the right of I, resp. to the left of I. Then either $I_r\subset J$ or $I_l\subset J$. Let assume that $I_r\subset J$ and leave the other case to the reader.
For every $i\in I$, we have

$$i+\ln 2\sqrt{n}/\kappa\in I_r.$$

(To simplify notations, we assume that $\ln 2\sqrt{n}/\kappa$ is an natural number). Let $j:=i+\ln 2\sqrt{n}/\kappa$. By assumption 3.17, we find that

$$P(W=j)\geq P(W=i)\cdot\left(1-\frac{\kappa}{\sqrt{n}}\right)^{\ln 2\sqrt{n}/\kappa} \tag{3.19}$$

The last inequality above yields

$$P(W \in I_r) \geq P(W \in I) \cdot \left(1 - \frac{\kappa}{\sqrt{n}}\right)^{\ln 2\sqrt{n}/\kappa}.$$

Hence,

$$P(W \notin I) \geq P(W \in I) \cdot \left(1 - \frac{\kappa}{\sqrt{n}}\right)^{\ln 2\sqrt{n}/\kappa} \tag{3.20}$$

Note that

$$\lim_{n\to\infty} \left(1 - \frac{\kappa}{\sqrt{n}}\right)^{\ln 2\sqrt{n}/\kappa} = \frac{1}{2},$$

and hence for n large enough

$$\left(1 - \frac{\kappa}{\sqrt{n}}\right)^{\ln 2\sqrt{n}/\kappa} > \frac{1}{3}.$$

The last inequality together with 3.20, yields

$$P(W \notin I) \geq \frac{P(W \in I)}{3},$$

from which it follows

$$P(W \notin I) \geq \frac{1}{4}. \tag{3.21}$$

We have that

$$\mathrm{Var}[W] = \sum_{i\in J} p_i \cdot (i - E[W])^2 \geq \sum_{i\in J/I} p_i \cdot (i - E[W])^2, \tag{3.22}$$

where $p_i := P(W = i)$.
Note that for every $i \notin I$, we have

$$(i - E[W])^2 \geq \left(\frac{\sqrt{n}\ln 2}{2\kappa}\right)^2.$$

Hence, the right side of inequality 3.22, is larger than

$$P(W \notin I) \cdot n\frac{(\ln 2)^2}{4\kappa^2}.$$

Using this with inequality 3.21 in 3.22, we obtain

$$\mathrm{Var}[W] \geq n \cdot \frac{(\ln 2)^2}{16\kappa^2}.$$

■

The next lemma will be useful:

Lemma 3.4 *There exists a constant $c_M > 0$ not depending on $\vec{M}$ or n such that*

$$\mathrm{Var}[M_1|\vec{M}, \vec{N} \in I^n] \geq c_M \cdot n \tag{3.23}$$

for all $\vec{M}$ such that $P(\vec{M}, \vec{N} \in I^n) > 0$.

Proof. Assume that $\vec{n} = (n_1, n_2, n_4, n_5) \in I^n$. Let $m_1 \in \mathbb{N}$ and $\vec{m} \in H_a \cup H_b$, be such that

$$\vec{n} = \vec{m} + m_1\vec{e}.$$

(Note that the last equality uniquely determines $\vec{m}$ and m_1).
Let I_1 be the integer interval such that

$$(\mathbb{N} \cdot \vec{e} + \vec{m}) \cap I^n = I_1 \cdot \vec{e} + \vec{m}.$$

We also write $I(\vec{m})$ to indicate that I_1 only depends on $\vec{m}$. Note that the condition

$$\vec{M} = \vec{m}, \vec{N} \in I^n$$

can now be rewritten as

$$\vec{M} = \vec{m}, M_1 \in I_1.$$

Equation 3.3 can be rewritten as:

$$\frac{P(\vec{M} = \vec{m}, M_1 = m_1 + 1)}{P(\vec{M} = \vec{m}, M_1 = m_1)} = \frac{(n_2 + 1)(n_4 + 1)}{n_1 n_5} \tag{3.24}$$

and hence

$$\frac{P(M_1 = m_1 + 1 | \vec{M} = \vec{m}, M_1 \in I_1)}{P(M_1 = m_1 | \vec{M} = \vec{m}, M_1 \in I_1)} = \frac{(n_2 + 1)(n_4 + 1)}{n_1 n_5} \tag{3.25}$$

We have that there exists $a, b, c, d > 0$ not depending on n, such that the following four inequalities hold

$$|\mu_1 - r_1| < a\sqrt{n},$$
$$|\mu_2 - r_2| < b\sqrt{n},$$
$$|\mu_4 - r_4| < c\sqrt{n},$$
$$|\mu_5 - r_5| < d\sqrt{n},$$

for any $\vec{r} = (r_1, r_2, r_3, r_4) \in I^n$.
Hence there exist $a^*, b^*, c^*, d^* > 0$, such that

$$\frac{(n_2 + 1)(n_4 + 1)}{n_1 n_5} = \frac{(\mu_2 + b^*\sqrt{n})(\mu_4 + c^*\sqrt{n})}{(\mu_1 + a^*\sqrt{n})(\mu_5 + d^*\sqrt{n})}, \tag{3.26}$$

and such that $|a^*| < a$, $|b^*| < b$, $|c^*| < c$ and $|d^*| < d$. Note that

$$\mu_i = \frac{n}{2^i},$$

which implies that

$$\frac{(\mu_2 + b^*\sqrt{n})(\mu_4 + c^*\sqrt{n})}{(\mu_1 + a^*\sqrt{n})(\mu_5 + d^*\sqrt{n})} = \frac{(0.5^2 + b^*/\sqrt{n})(0.5^4 + c^*/\sqrt{n})}{(0.5^1 + a^*/\sqrt{n})(0.5^5 + d^*/\sqrt{n})} \tag{3.27}$$

Consider the map

$$f:(x,y,z,w)\mapsto \frac{(0.5^2+x)(0.5^4+y)}{(0.5^1+z)(0.5^4+w)}.$$

we have $f(0,0,0,0)=1$. Note that the maps $f(x,y,z,w)$ and $1/f(x,y,z,w)$ are both continuously differentiable in an open neighborhood of $\vec{0}=(0,0,0,0)$. This implies that there exist a constant $\kappa>0$ not depending on n, such that for all n large enough, we have

$$\frac{1}{1-\kappa\sqrt{n}}\geq \frac{(0.5^2+b^*/\sqrt{n})(0.5^4+c^*/\sqrt{n})}{(0.5^1+a^*/\sqrt{n})(0.5^5+d^*/\sqrt{n})}\geq 1-\kappa\sqrt{n} \tag{3.28}$$

which holds for every a^*,b^*,c^*,d^* such that $|a^*|\leq a$, $|b^*|\leq b$, $|c^*|\leq c$ and $|d^*|\leq d|$. Combining, 3.25, 3.26, 3.27 and 3.28, we find that

$$\frac{P(M_1=m_1+1|\vec{M}=\vec{m},M_1\in I_1)}{P(M_1=m_1|\vec{M}=\vec{m},M_1\in I_1)}\geq 1-\kappa\sqrt{n}$$

and

$$\frac{P(M_1=m_1|\vec{M}=\vec{m},M_1\in I_1)}{P(M_1=m_1+1|\vec{M}=\vec{m},M_1\in I_1)}\geq 1-\kappa\sqrt{n}.$$

Note that the two last inequalities above are the conditions 3.16 and 3.17 in lemma 3.3. For this we take for the random variable W of lemma 3.3, the random variable M_1 conditional on $\vec{M}=\vec{m}$ and $M_1\in I_1$. In order to apply lemma 3.3, we also need to verify that the interval $I(\vec{m})$ has diameter at least $3\kappa\ln 2\sqrt{n}$. For this note that by choosing the constant $C>0$ in the definition 3.6 of I^n large enough (but not depending on n), we get that the diameter of $I(\vec{m})$ is larger than $3\kappa\ln 2\sqrt{n}$ for every $\vec{m}$, such that

$$P(\vec{M}=\vec{m},\vec{N}\in I^n)>0.$$

We can now apply lemma 3.3, and obtain that

$$\mathrm{Var}[M_1|\vec{M}=\vec{m},M_1\in I_1]=\mathrm{Var}[M_1|\vec{M}=\vec{m},\vec{N}\in I^n]\geq n\cdot\frac{(\ln 2)^2}{16\kappa^2}.$$

■

We can now apply inequality 3.23 to 3.15. We find that conditional on $\vec{N}\in I^n$ and when E^n_{slope} holds,

$$\mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\geq n\cdot\left(0.01^2\left(1-\frac{2n^{0.1}}{0.01\sqrt{c_M\cdot n}}\right)c_M\right). \tag{3.29}$$

Using inequality 3.29, we have that inequality 3.9 becomes:

$$E\left[\ \mathrm{Var}[L(M_1,\vec{M})|L(.),\vec{M}]\ \right]\geq \tag{3.30}$$

$$\geq n\cdot\left(c_M\ 0.01^2\left(1-\frac{2}{0.01\sqrt{c_M}n^{0.4}}\right)\right)\cdot P(\vec{N}\in I^n)\cdot P(E^n_{\text{slope}}) \tag{3.31}$$

With inequality 3.8 and equality 3.4, we then find

$$\mathrm{Var}[L_n] \geq n \cdot \left(c_M\ 0.01^2 \left(1 - \frac{2}{0.01\sqrt{c_M} n^{0.4}} \right) \right) \cdot P(\vec{N} \in I^n) \cdot P(E^n_{\text{slope}}) \tag{3.32}$$

To finished the proof of theorem 2.1 it suffices to prove that $P(\vec{N} \in I^n)$ and $P(E^n_{\text{slope}})$ are bounded below by a positive number when n goes to infinity. This is the content of the two next lemmas.

Lemma 3.5 *We have that*

$$P(\vec{N} \in I^n) \geq \frac{1}{5}.$$

Proof. Let Z_i be the indicator function which is equal to one if

$$X_i = 1, X_{i+1} = 0, X_{i+2} = 1.$$

and $Z_i = 0$ otherwise. In other words, $X_i = 1$ if there is a block of one zero in X starting at the point i. We have that

$$N_1 = \sum_{i=1}^{n-3} Z_i,$$

and hence

$$\mathrm{Var}[N_1] = E[(\sum_{i=1}^{n-3}(Z_i - E[Z_i]))^2] = \sum_{i,j} E[(Z_i - E[Z_i])(Z_j - E[Z_j])]. \tag{3.33}$$

When $|i-j| > 2$, then Z_i and Z_j are independent of each other and:

$$E[(Z_i - E[Z_i])(Z_j - E[Z_j])] = 0.$$

This implies that in the sum on the very right side of 3.33, there at most $5n$ terms which are different from zero. By Cauchy-Schwarz, we have

$$E[(Z_i - E[Z_i])(Z_j - E[Z_j])] \leq \sqrt{E[(Z_i - E[Z_i])^2] \cdot E[(Z_j - E[Z_j])^2]} \leq \mathrm{Var}[Z_1].$$

For a Bernoulli random variable the variance is always, less than 1/2. Hence, for all $i, j \in [0, n-3]$ we have

$$E[(Z_i - E[Z_i])(Z_j - E[Z_j])] \leq \frac{1}{4} \tag{3.34}$$

since there are less than $5n$ non-zero terms in the sum on the right side of 3.33, we obtain with 3.34 that:

$$\mathrm{Var}[N_1] \leq \frac{5n}{4} \tag{3.35}$$

Similarly we obtain

$$\mathrm{Var}[N_2] \leq \frac{5n}{4}, \quad \mathrm{Var}[N_4] \leq \frac{9n}{4}, \quad \mathrm{Var}[N_5] \leq \frac{11n}{4}, \tag{3.36}$$

In 3.7, we have chosen I^n such that

$$\bigcap_{i=1,2,4,5} \left\{ N_i \in [\ \mu_i - \sqrt{55n/4}, \mu_i + \sqrt{55n/4}\ ] \right\} \quad \subset \quad \left\{ \vec{N} \in I^n \ \right\}$$

and hence

$$P(\vec{N} \notin I^n) \leq \sum_{i=1,2,4,5} P(N_i \notin [\ \mu_i - \sqrt{55n/4}, \mu_i + \sqrt{55n/4}\ ]\ ) \tag{3.37}$$

Chebycheff inequality yiels that

$$P(\ N_i \notin [\mu_i - \sqrt{55n/4}, \mu_i + \sqrt{55n/4}] \leq \frac{\mathrm{Var}[N_i]}{55n/4},$$

and hence together with 3.36 and 3.37, this yields:

$$P(\vec{N} \notin I^n) \leq \sum_{i=1,2,4,5} \frac{11n}{4} \cdot \frac{4}{55n} = \frac{4}{5}.$$

This finishes the proof. ■

The next lemma will be useful:

Lemma 3.6 *Recall that μ_i designates the expected number of blocks of zeros of length i in X.*
For all n and all i we have that

$$\left| \mu_i - n \cdot \left(\frac{1}{2}\right)^{i+2} \right| \leq i \tag{3.38}$$

Proof. Let R^i_j be the Bernoulli random variable which indicates if there is a block of zeros of length i starting at the point j in X. In other words, $R^i_j = 1$ iff

$$X_j = X_{j+1} = \ldots = X_{j+i-1} = 0 \ \text{ and } \ X_{j-1} = X_{j+i} = 1$$

for $j \in [2, n-i]$.
For $j = 1$ we have $R^i_j = 1$ iff

$$X_1 = X_2 = \ldots = X_i = 0 \ \text{ and } \ X_{i+1} = 1.$$

For $j = n - i + 1$, we have that $R^i_j = 1$ iff

$$X_{n-i+1} = X_{n-i+2} = \ldots = X_n = 0 \ \text{ and } \ X_{n-i} = 1.$$

We find that

$$N_i = \sum_{j=1}^{n-i+1} R^i_j$$

and hence

$$\mu_i = E[N_i] = \sum_{j=1}^{n-i+1} E[R^i_j]. \tag{3.39}$$

Note that for $j \in [2, n-i]$ we have that $E[R^i_j] = (1/2)^{i+2}$.
We also have that $E[R^i_1] = E[R^i_{n-i+1}] = (1/2)^{i+1}$.
This with 3.39, yields

$$\mu_i = (n-i) \cdot (1/2)^{i+2} + 2 \cdot (1/2)^{i+1}. \tag{3.40}$$

The last equality directly implies 3.38. ∎

Our next lemma is a related to local central limit theorem for multidimensional regenerative processes:

Lemma 3.7 *There exist a constant $k_1 > 0$ such that for all all n, we have: for all $\vec{n} \in I^n$, the inequality*

$$P(\vec{N} = \vec{n}) \geq n^{-k_1} \tag{3.41}$$

holds.

Proof. Let T_i be the end of the i-th block in the infinite sequence $X_1, X_2, X_3, \ldots$. To simplify notation, we assume that the finite sequence $X = X_1 X_2 \ldots X_n$ got extended to an infinite sequence of iid Bernoulli random variables with parameter 1/2. In this way, T_i is well defined even when $T_i > n$. (Note that here we consider all the blocks, and not just the blocks of zeros). Let Z_i denote the length of the i-th block of X. Hence, we have that $Z_i = T_i - T_{i-1}$ and

$$T_i = Z_1 + Z_2 + \ldots + Z_i.$$

Let us give a numerical example. Take $X = 00111001\ldots$. Then we find $T_1 = 3$, $T_2 = 5$ and $T_3 = 8$.

Note that $T_1, T_2, \ldots$ are the arrival times of a renewal process. The interarrival times $Z_1, Z_2, \ldots$ are iid geometric random variables with parameter 2 and hence $E[Z_i] = 2$.
We assume that $n/4$ is an integer in order to simplify notation. There are $n/2$ blocks in X iff $T_{n/2} = n$.
The event that the blocks of ones cover half the text X can be described by the equality

$$\sum_{i=1}^{n} X_i = \frac{n}{2}. \tag{3.42}$$

Let $\vec{n} = (n_1, n_2, n_4, n_5) \in I^n$. We have that

$$P(\vec{N} = \vec{n}) \geq P(\vec{N} = \vec{n}, \sum_{i=1}^{n} X_i = \frac{n}{2}, T_{n/2} = n). \tag{3.43}$$

Let V_i^j be the indicator variable which is equal to one iff the i-th block of zeros in X had length j and $V_i^j = 0$ otherwise. Let

$$\vec{V_i} := (V_i^1, V_i^2, V_i^4, V_i^5).$$

Note that $\vec{V}_1, \vec{V}_2, \ldots$ constitute a sequence of i.i.d. random vectors.
The event

$$\{\vec{N} = \vec{n}, \sum_{i=1}^{n} X_i = \frac{n}{2}, T_{n/2} = n\}$$

can be interpreted as the event that there are $n/2$ blocks in X and that the blocks of ones cover half the text and that $\vec{N} = \vec{n}$. Its probability can be calculate as follows: First calculate the contribution due to the fact that the blocks of ones cover half the text. Half of all blocks are blocks of ones. Hence we have exactly $n/4$ blocks of ones. These blocks are i.i.d. with geometric distribution with parameter 1/2. This gives us a factor

$$P(Z_1 + Z_2 + \ldots + Z_{n/4} = n/2) \tag{3.44}$$

Second, we compute the probability that among the first $n/4$ blocks of zeros we have the right number of blocks of length 1, 2, 4 and 5. We find

$$P(\vec{V}_1 + \ldots + \vec{V}_{n/4} = \vec{n}). \tag{3.45}$$

Finally we calculate the probability that the remaining $n/4 - n_1 - n_2 - n_4 - n_5$ blocks of zeros, cover up a total length of $n/2 - n_1 - 2n_2 - 4n_4 - 5n_5$. This yields the probability

$$P(W_1 + W_2 + \ldots + W_{n^*} = n/2 - n_1 - 2n_2 - 4n_4 - 5n_5), \tag{3.46}$$

where $n^* := n/4 - n_1 - n_2 - n_4 - n_5$ and $W_1, W_2, \ldots$ is a sequence of iid random variables with distribution

$$\mathcal{L}(Z_i | Z_i \notin \{1, 2, 4, 5\}).$$

Summarizing, we find

$$P(\vec{N} = \vec{n}, \sum_{i=1}^{n} X_i = \frac{n}{2}, T_{n/2} = n) =$$

$$P(Z_1 + \ldots + Z_{n/4} = n/2) \cdot P(\vec{V}_1 + \ldots + \vec{V}_{n/4} = \vec{n}) \cdot P(W_1 + \ldots + W_{n^*} = n/2 - \bar{n}),$$

where $\bar{n} = n_1 + 2n_2 + 4n_4 + 5n_5$ and hence with 3.43:

$$P(\vec{N} = \vec{n}) \geq$$

$$P(Z_1 + \ldots + Z_{n/4} = n/2) \cdot P(\vec{V}_1 + \ldots + \vec{V}_{n/4} = \vec{n}) \cdot P(W_1 + \ldots + W_{n^*} = n/2 - \bar{n}).$$

We have that $(n/4) \cdot E[Z_1] = n/2$. Hence by the local central limit theorem, we find that there exists $k_2 > 0$ not depending on n such that

$$P(Z_1 + \ldots + Z_{n/4} = n/2) \geq k_2 \cdot n^{-0.5} \tag{3.47}$$

for all n.
Note that

$$E[V_1^i] = \left(\frac{1}{2}\right)^i. \tag{3.48}$$

The last equality together with inequality 3.38, yields:

$$|\vec{\mu} - (n/4) \cdot E[\vec{V}_1]| \leq 1 + 2 + 4 + 5 = 12, \tag{3.49}$$

where $|.|$ denotes the norm in $\mathbb{R}^4$ for which $|(x, y, z, w)| = |x| + |y| + |z| + |w|$.
We have that

$$|\vec{n} - (n/4) \cdot E[\vec{V}_1]| \leq |\vec{n} - \vec{\mu}| + |\vec{\mu} - (n/4) \cdot E[\vec{V}_1]|. \tag{3.50}$$

Since, $\vec{n} \in I^n$ we have that there exist a constant $K > 0$ not depending on n or $\vec{n} \in I^n$ such that

$$|\vec{n} - \vec{\mu}| \leq K \cdot \sqrt{n} \tag{3.51}$$

Combining 3.49, 3.50 and 3.51 yields

$$|\vec{n} - (n/4) \cdot E[\vec{V}_1]| \leq K \cdot \sqrt{N} + 12$$

which hold for every n and every $\vec{n} \in I^n$. From the last inequality above, with the help of the local central limit theorem, we obtain that there exists $k_3 > 0$ not depending on n or $\vec{n} \in I^n$ such that

$$P(\vec{V}_1 + \ldots + \vec{V}_{n/4} = \vec{n}) \geq k_3 \cdot n^{-2}. \tag{3.52}$$

Next, we want to prove that

$$|n^* E[W_1] - n/2 + n_1 + n_2 + n_4 + n_5| \tag{3.53}$$

is of order $\sqrt{n}$. Let

$$u_i := \frac{n}{4} \left(\frac{1}{2}\right)^i.$$

We know there is a constant $K >$ not depending on n or $\vec{n} \in I^n$, such that

$$|\mu - \vec{n}| \leq K\sqrt{n}.$$

From this and and from 3.38 it follows that there exist $K_2 > 0$ such that the difference between 3.53 and

$$|(n/4 - u_1 - u_2 - u_4 - u_5)E[W_1] - n/2 + u_1 + u_2 + u_4 + u_5| \tag{3.54}$$

is less than $K_2 \cdot \sqrt{n}$.
Now consider a sequence of $n/4$ iid geometric random variables with parameter $1/2$. The expectation of the sum of these random variables is:

$$E[T_{n/4}] = \frac{n}{4} \cdot E[Z_1] = \frac{n}{2}.$$

Let W denote the sub-sum obtained by only taking the terms not equal to 1,2,4 or 5. Let $\bar{W}$ denote the sub-sum obtained by taking only those terms in the sum which are equal to 1,2,4 or 5.
Let b denote the total number of random variables among our collection of $n/4$, that take a value equal to 1,2,4 or 5.

Note that the probability for one variable to take on value i is equal to $(1/2)^i$. Hence, the expected number of variables among our set of $n/4$ which take on value i is equal to $(n/4) \cdot (1/2)^i = u_i$. We find

$$E[T_{n/4}]] = E[W + \bar{W}] = E[W] + E[\bar{W}],$$

but

$$E[W] = E[W_1] \cdot E[(n/4) - b] = E[W_1] \cdot ((n/4) - u_1 - u_2 - u_4 - u_5)$$

and

$$E[\bar{W}] = u_1 + 2u_2 + 4u_4 + 5u_5.$$

Combining the last three inequalities yields:

$$n/2 = E[T_{n/4}]] = E[W_1] \cdot ((n/4) - u_1 - u_2 - u_4 - u_5) + u_1 + 2u_2 + 4u_4 + 5u_5,$$

and hence

$$0 = E[W_1] \cdot ((n/4) - u_1 - u_2 - u_4 - u_5) - (n/2) + u_1 + 2u_2 + 4u_4 + 5u_5 \quad (3.55)$$

Hence expression 3.54 is equal to zero and thus

$$|n^* E[W_1] - n/2 + n_1 + n_2 + n_4 + n_5| \leq K_2 \cdot \sqrt{n}.$$

The above inequality used with the local central limit theorem yields that there exists $k_4 > 0$ not depending on n or $\vec{n} \in I^n$ such that

$$P(W_1 + W_2 + \ldots + W_{n^*} = n/2 - n_1 - 2n_2 - 4n_4 - 5n_5) \geq k_4 \cdot n^{1/2} \quad (3.56)$$

Using inequality 3.47, 3.52 and 3.56 implies 3.41. ■

The next lemma is proved assuming that theorem 2.2 holds. (For the proof of theorem 2.2 see subsection 5.2).

Lemma 3.8 *There exists $s > 0$ not depending on n such that if $s(1,1) \geq s$, then*

$$P(E^n_{\text{slope}}) \to 1,$$

as $n \to \infty$.

Proof. Let $\vec{m} \in H_a \cup H_b$, be such that

$$P(\vec{M} = \vec{m}, \vec{N} \in I^n) > 0. \quad (3.57)$$

Let $E^n_{\vec{m}}$ be the event that E^n_{slope} holds on the subset $I(\vec{m}) \cdot \vec{e} + \vec{m}$. (The set $I(\vec{m})$ got defined in the proof of lemma 3.4). More precisely, $E^n_{\vec{m}}$ is the event that for all $i, j \in I(\vec{m})$, such that $j - i \geq n^{0.1}$, we have that

$$L(\vec{m} + j\vec{e}) - L(\vec{m} + i\vec{e}) \geq 0.01 \cdot |j - i|.$$

We find that

$$E^n_{\text{slope}} = \bigcap_{\vec{m}} E^n_{\vec{m}}, \tag{3.58}$$

where the intersection on the right side of the last equality above is taken over all $\vec{m} \in H_a \cup H_b$, such that 3.57 holds. Note that there exists a constant c_I (not depending on n) such that for all n there are less than $c_I \cdot n^2$ points in the set I^n. Hence there are also less than $c_I \cdot n^2$ vectors $\vec{m} \in H_a \cup H_b$ satisfying 3.57. It follows that in the intersection on the right side of 3.58, there are less than $c_I \cdot n^2$ terms. Now, 3.58 implies that

$$P(E^{nc}_{\text{slope}}) \leq \sum_{\vec{m}} P(E^{nc}_{\vec{m}}), \tag{3.59}$$

where the sum on the right side of the last inequality is taken over all $\vec{m} \in H_a \cup H_b$, such 3.57 holds. There are less than $c_I \cdot n^2$ terms in the sum on the right side of 3.59. Hence to prove this lemma, we only need a negative exponential upper-bound for $P(E^{nc}_{\vec{m}})$ which does not depend on $\vec{m}$.
Let A^n_i denote the event "that A^n holds for $X(\vec{m} + i\vec{e})$". More precisely, A^n_i is the event that the following two conditions hold:

$$P(L(\vec{m} + (i+1)\vec{e}) - L(\vec{m} + i\vec{e}) = 1 | X(\vec{m} + i\vec{e}), Y) \geq \frac{31}{32} \cdot \frac{1}{4} - \epsilon_1, \tag{3.60}$$

$$P(L(\vec{m} + (i+1)\vec{e}) - L(\vec{m} + i\vec{e}) = -1 | X(\vec{m} + i\vec{e}), Y) \leq \frac{1}{32} + \epsilon_1 \tag{3.61}$$

Note that the difference between $L(\vec{m} + (i+1)\vec{e})$ and $L(\vec{m} + (i+1)\vec{e})$ is at most one. We assume that the constant $\epsilon_1 \leq 7/32$. The inequalities 3.60 and 3.61 give then that for every $(x, y) \in A^n_i$, we have

$$E[\ L(\vec{m} + (i+1)\vec{e}) - L(\vec{m} + i\vec{e}) = 1 | X(\vec{m} + i\vec{e}) = x, Y = y) \geq \frac{1}{2} \tag{3.62}$$

From a positive bias like in 3.62, one can hope to prove that the event $E^n_{\vec{m}}$ holds with probability one minus and stretched exponential small quantity. The only problem is that inequality 3.62 holds only for $(x, y) \in A^n_i$. If we condition on $\cap_{i \in I(\vec{m})} A^n_i$, we introduce complicated dependencies so that we can no longer use large deviation results for martingales. The trick is to introduce help-variables Y_i. When, A^n_i holds let

$$Y_i := L(\vec{m} + (i+1)\vec{e}) - L(\vec{m} + i\vec{e}),$$

otherwise let $Y_i := 1$. We have that Y_i is $\sigma(Y, X(\vec{m} + j\vec{e}) | j \leq i)$-measurable. Also, because of 3.62, we have that almost surely

$$E[Y_i | Y, X(\vec{m} + j\vec{e})] \geq \frac{1}{2}. \tag{3.63}$$

Let $E^n_{\vec{m},Y}$ be the event that for all $i, j \in I(\vec{m})$, such that $j - i \geq n^{0.1}$, we have that

$$\sum_{k=i}^{j} Y_k \geq 0.01 \cdot |j - i|.$$

Note that when $\cap_{i\in I(\vec{m})}A_i^n$ holds, then the events $E_{\vec{m},Y}^n$ and $E_{\vec{m}}^n$ are identical. Hence,

$$\cap_{i\in I(\vec{m})}A_i^n \ \cap \ E_{\vec{m},Y}^n \subset E_{\vec{m}}^n,$$

and hence

$$P(E_{\vec{m}}^{nc}) \leq P(E_{\vec{m},Y}^{nc}) + \sum_{i\in I(\vec{m})} P(A_i^{nc}). \tag{3.64}$$

Note that $X(\vec{m}+i\vec{e})$ has distribution $\mathcal{L}(X|\vec{N}=\vec{m}+i\vec{e})$. This implies that

$$P(A_i^{nc}) = \frac{P(A^{nc}\cap\{\ \vec{N}=\vec{m}+i\vec{e}\ \})}{P(\vec{N}=\vec{m}+i\vec{e})} \leq \frac{P(A^{nc})}{P(\vec{N}=\vec{m}+i\vec{e})}$$

Using the last inequality with 3.41, we obtain

$$P(A_i^{nc}) \leq P(A^{nc})\cdot n^{k_1}.$$

The last inequality implies:

$$\sum_{i\in I(\vec{m})} P(A_i^{nc}) \leq c_I\cdot n^{2+k_1}P(A^{nc}). \tag{3.65}$$

Classical large deviations and inequality (3.63) show that

$$P(E_{\vec{m},Y}^{nc}) \leq e^{-k_Y\cdot n^{0.1}}, \tag{3.66}$$

where $k_Y>0$ is a constant not depending on n. Inequality 3.64 with 3.65 together imply

$$P(E_{\vec{m}}^{nc}) \leq e^{-k_Y\cdot n^{0.1}} + c_I\cdot n^{2+k_1}P(A^{nc}). \tag{3.67}$$

We can now plug in inequality 3.32 into 3.67 and obtain

$$P(E_{\vec{m}}^{nc}) \leq e^{-k_Y\cdot n^{0.1}} + c_I\cdot n^{2+k_1}e^{-c_1 n}.$$

The upper-bound in the last inequality above is exponentially small in $n^{0.1}$. Hence by 3.59, we get that $P(E_{\text{slope}}^{nc})$ must also be exponentially small in $n^{0.1}$.

■

4 Combinatorics

We already mentioned that $P(X_i=1)=1/5$. Let $\epsilon>0$ designate a small quantity not depending on n.

Let B_0^n designate the event that in both X and Y there are about $0.5n$ ones. More precisely, B_0^n is the event that

$$|\sum_{i=1}^n X_i - 0.5n| \leq \frac{\epsilon n}{16}$$

and

$$|\sum_{i=1}^{n} Y_i - 0.5n| \leq \frac{\epsilon n}{16}$$

both hold. Let B_1^n be the event that any optimal alignment of X and Y contains at least $0.5n - \epsilon n/8$ pairs of aligned ones.

Lemma 4.1 *Assume that $s(1,1)$ and $\epsilon > 0$ satisfy 4.3. Then, we have that*

$$B_0^n \subset B_1^n \tag{4.1}$$

for all n.

Proof. Let v_1 be the alignment of X and Y which aligns only ones and as many as possible. Let S_1 denote the score obtained by aligning X with Y via v. When, B_0^n holds, we obtain

$$S_1 \geq s(1,1) \cdot (0.5n - \epsilon n/16),$$

and therefore

$$L_n \geq s(1,1) \cdot (0.5n - \epsilon n/16). \tag{4.2}$$

We will assume that

$$s(1,1) > \frac{16}{\epsilon} \tag{4.3}$$

Inequality 4.3 implies

$$(0.5n - \epsilon n/16) \cdot s(1,1) > (0.5n - \epsilon n/8) \cdot s(1,1) + n,$$

which applied to 4.2 gives

$$L_n > (0.5n - \epsilon n/8) \cdot s(1,1) + n. \tag{4.4}$$

Recall that $s(0,0) = 1$. Hence, if the texts X and Y consist only of zeros the maximum score would be equal to n. This also implies that n is an upper bound for the contribution made by the aligned zeros to the score of any alignment. Assume now that v is an alignment which aligns no more that $0.5n - \epsilon n/8$ pairs of ones. Let S_v denote the score of v. When B_0^n holds, using the bound for the contributions of the zeros in the score, we find that

$$S_v \leq (0.5n - \epsilon n/8) \cdot s(1,1) + n. \tag{4.5}$$

Together, 4.4 and 4.5 imply that v is not an optimal alignment. This implies that when B_0^n holds, than any optimal alignment contains at least $0.5n - \epsilon n/8$ pairs of aligned ones. In other words, B_0^n implies B_1^n, when 4.3 holds. ■

Recall that we take the score $s(1,1)$ high. This insures, that typically there is a large proportion of the total number of ones, which get matched with a one by the optimal alignment. We introduce a special notation for alignments, which is convenient to describe alignments which align most ones with ones. Let us start with a numerical example:

Take the finite sequence of pairs of natural numbers

$$(0,0),(0,1),(1,0).$$

According to our notations, this sequence represents an alignment which does the following:
-First $(0,0)$ means that we align the first one of X with the first one of Y without skipping any one.
-The second pair $(0,1)$ indicates that after the first pair of aligned ones, we skip one one in the Y sequence and no one in the X sequence.
-The third pair $(1,0)$ means that after the second pair of aligned ones, we skip a one in X and zero one in Y.
Take for example the sequences $X = 0101011$ and $Y = 0001111$. The alignment $v = ((0,0),(0,1),(1,0))$ is then equal to:

0			1	0	1	0	1	1
0	0	0	1	1	1			1

Recall that the score for aligning a zero with a one is zero: $s(0,1) = 0$.

Let V^k designate the set of alignments which align exactly k pairs of of ones with each other and such that there is a proportion of less than $\epsilon/2$ ones not belonging to pairs of aligned ones. (The ones which are not aligned with ones are counted up to the last pair of aligned ones.) Hence, with our representation of alignments of pairs of ones as sequences of couples of natural numbers we find

$$V^k := \{(v_1, v_2, \ldots, v_k) | v_1, v_2, \ldots, v_k \in \mathbb{N} \times \mathbb{N}, |v_1| + |v_2| + \ldots + |v_k| \leq \epsilon k/2\},$$

where if $v = (a,b) \in \mathbb{N}^2$, we define the norm $|v| := a + b$. Let

$$V := \cup_{k \geq p^* n} V^k$$

where $p^* := 0.5 - \epsilon/8$.
Let B_2^n be the event that any optimal alignment of X with Y is contained in V. In other words, B_2^n holds when for every alignment such that $S_v = L_n$, we have $v \in V$. (Here S_v designates the score obtained by aligning X with Y and using for this the alignment v).
The next lemma shows that B_0^n and B_1^n together imply B_n^2:

Lemma 4.2 *When $\epsilon > 0$ satisfies 4.8, we have that*

$$B_0^n \cap B_1^n \subset B_2^n \tag{4.6}$$

for all n.

Proof. Let v be an alignment. We say a one is *matched* by v if it gets aligned with another one by v. When it is clear from the context which alignment v we are talking about, we simply say that a one is matched. If B_0^n and B_1^n both hold, there are at most $(\epsilon n/16) + (\epsilon n/8)$ non-matched ones in each text X and Y for any optimal alignment v. On the other hand, B_1^n insures that at least $0.5n - \epsilon n/8$ ones are matched in each text with a one from the other text by any optimal alignment v.

This insures that the proportion of non-matched ones by total number of matched ones is smaller/equal than:

$$\frac{\frac{\epsilon n}{16}+\frac{\epsilon n}{8}}{0.5n-\epsilon n/8}=\frac{3\epsilon}{8}\left(\frac{1}{1-\epsilon/4}\right). \tag{4.7}$$

Note that

$$\lim_{\epsilon\to 0}\frac{1}{1-\epsilon/4}=1,$$

and hence by choosing $\epsilon>0$ small enough, we get

$$\frac{3\epsilon}{8}\left(\frac{1}{1-\epsilon/4}\right)\leq\frac{\epsilon}{2}. \tag{4.8}$$

We assume henceforth that inequality 4.8 holds. This implies that the optimal alignment has a proportion of non-matched ones to matched ones smaller than $\epsilon/2$. Adding to this that by B_1^n there are at least $0.5n-\epsilon n/8$ pairs of aligned ones, we get that any optimal alignment is in V. Hence, B_2^n holds. We have just proved that B_0^n and B_1^n together imply B_2^n, when $\epsilon>0$ is small enough. ■

For $v\in V^k$, let $(\pi_v(i),\nu_v(i))$ be the indexes of the i-th pair of ones aligned by v. Hence, if $\pi_v(i)=j$ and $\nu_v(i)=k$, then all of the following holds:
-X_j gets aligned with Y_k.
-$X_j=Y_k=1$
-The i-pair of aligned ones (by v) is X_j and Y_k.

In the last numerical example above we have that $\pi_v(1)=2$, $\pi_v(2)=4$ and $\pi_v(3)=7$. Furthermore, $\nu_v(1)=4$, $\nu_v(2)=6$ and $\nu_v(3)=7$.

Let v designate an alignment. Recall that a one that gets aligned by v to another one (instead of to a gap or a zero), is called *matched* by v.
Let $N_5(v)$ denote in the alignment v, the total number of subsequent pairs of ones in X satisfying all of the following conditions:
-The ones are both matched, that is aligned by v with a one from Y.
-Between the pair of ones in the X text there are only zeros. More precisely, we require that there is a block of five zeros, between the pair of ones in the X text.
-The pair of ones in the Y text with which our pair from X is aligned, should contain only zeros in between them.
More precisely: for $v\in V^k$, we define:

$$N_5(v):=|\{i<k|v_i=(0,0),\pi_v(i+1)-\pi_v(i)=6\}|$$

where $v:=(v_1,v_2,\ldots,v_k)$.
Let $N_{5<}(v)$ denote in the alignment v, the total number of subsequent pairs of ones in X satisfying all of the following conditions:
-The ones are both matched, that is aligned by v with a one from Y.
-Between the pair of ones in the X text there are only zeros and we require that there exactly five zeros.

-The pair of ones in the Y text with which our pair from X is aligned, should contain only zeros in between them and contained strictly less than 5.
More precisely:

$$N_{5<}(v) := |\{i < k | v_i = (0,0), \pi(i+1) - \pi(i) = 6, \nu_v(i+1) - \nu_v(i) < 6\}|.$$

Let C^n be the event that for all $v \in V$ which is an optimal alignment, we have that

$$\frac{N_{5<}(v)}{N_5(v)} \geq \frac{31}{32} - \epsilon_1/4.$$

Let B_3^n be the event that in the sequence X there are at least $((1/32) - \epsilon/16)n$ blocks of zeros of length five.
Let $p_5(v)$ be the conditional probability on X, that when we pick a block of five zeros at random in X, this block happens to satisfy the following two conditions:
1) The block is contained between two consecutive matched ones. (Matched by the alignment v.) This means that between the two matched ones there is our block of length five and nothing else.
2) The pair of ones in the Y-text to which the pair of consecutive ones are aligned by v contains only zeros in between them and strudel less than five of them.
In other words, $p_5(v)$ designates the conditional probability (conditional on X) that when we pick at random a block of zeros of length five in X, there exists $i \leq k$, such that the randomly selected block is equal to $[\pi_v(i-1)+1, \pi_v(i)-1]$ and all of the following holds:

$$\pi_v(i) - \pi_v(i-1) = 6,$$

$$\eta_v(i) - \eta_v(i-1) < 6$$

and $|v_i| = 0$, where

$$v = (v_1, v_2, \ldots, v_k) \in (\mathbb{N} \times \mathbb{N})^k.$$

Recall that the total number of blocks of five zeros in X is denoted by n_5. Furthermore, we select a block of length five with equal probability among all blocks of five zero in X. Hence, each block of five zeros in X has a conditional probability of $1/n_5$ to get selected. There, are $N_{5<}$ blocks of five zeros in X satisfying our conditions. Hence the conditional probability $p_5(v)$ is equal to:

$$p_5(v) := \frac{N_{5<}(v)}{n_5}.$$

Let us give a numerical example. Let $X = 101010101000001$ and let $Y = 101000110001$. Let v be the alignment

1	0	1	0			1	0	1	1	0	0	0	0	0	1
1	0	1	0	0	0	1			1	0	0	0			1

In X there is one block of five zeros. Hence, $n_5 = 1$. Furthermore this block is contained directly between matched ones and the corresponding ones in the Y-text contain between them only zeros. Hence, our block of five zeros counts towards $N_5(v)$. We have $N_5(v) = 1$. The block of five zeros in X is matched with a block with three zeros in Y, hence with a

block with strictly less zeros. Thus, the block with five zeros is also counted towards $N_{5<}$ and we find $N_{5<} = 1$.
The conditional probability p_5 is equal to 1.

Let E^n designate the event that

$$p_5(v) \geq \frac{31}{32} - \frac{\epsilon_1}{2}.$$

for every optimal alignment $v \in V$.

Lemma 4.3 *For $\epsilon > 0$ and $\epsilon_1 > 0$ satisfying inequality 4.14, we have*

$$B_0^n \cap B_1^n \cap B_3^n \cap C^n \subset E^n \tag{4.9}$$

for all n.

Proof. We have for any alignment $v \in V$, that

$$\frac{N_{5<}(v)}{n_5} = \frac{N_{5<}(v)}{N_5(v)} \cdot \frac{N_5(v)}{n_5}. \tag{4.10}$$

When, the event C^n holds, then, for every $v \in V$, we have

$$\frac{N_{5<}(v)}{N_5(v)} \geq \frac{31}{32} - \epsilon_1/4.$$

Combining the last inequality with equality 4.10, yields

$$\frac{N_{5<}(v)}{n_5} \geq \left(\frac{31}{32} - \epsilon_1/4\right) \cdot \frac{N_5(v)}{n_5}. \tag{4.11}$$

We have that B_0^n and B_1^n together imply B_2^n. Hence, when B_0^n and B_1^n both hold, then every optimal alignment is in V. Hence, inequality 4.11 also holds for every optimal alignment v.
If B_0^n and B_1^n both hold, there are at most $(\epsilon n/16) + (\epsilon n/8)$ non-matched ones in each text X and Y for any optimal alignment v. This also implies that the number of blocks of five zeros in X which do not satisfy the criteria to be counted towards $N_5(v)$ is at most $2[(\epsilon n/16) + (\epsilon n/8)]$, (for every optimal alignment v). This implies that for every optimal alignment v, when B_0^n and B_1^n both hold, then

$$N_5(v) \geq n_5 - 3n\epsilon/8.$$

The last inequality with inequality 4.11 implies

$$\frac{N_{5<}(v)}{n_5} \geq \left(\frac{31}{32} - \epsilon_1/4\right) \cdot \left(1 - \frac{3n\epsilon}{8n_5}\right), \tag{4.12}$$

for every optimal alignment v. When B_3^n holds, we have that

$$n_5 \geq ((1/32) - \epsilon/16)n,$$

and hence

$$1-\frac{3n\epsilon}{8n_5}\geq 1-\frac{3\epsilon}{1/4-\epsilon/2}.$$

Using the last inequality in inequality 4.12 gives:

$$\frac{N_{5<}(v)}{n_5}\geq\left(\frac{31}{32}-\epsilon_1/4\right)\cdot\left(1-\frac{3\epsilon}{1/4-\epsilon/2}\right). \tag{4.13}$$

Note that

$$\lim_{\epsilon\to 0}\left(1-\frac{3\epsilon}{1/4-\epsilon/2}\right)=1.$$

Hence for any $\epsilon_1>0$ fixed, if we chose $\epsilon>0$ small enough, (how small depends on ϵ_1), we get

$$\left(\frac{31}{32}-\epsilon_1/4\right)\cdot\left(1-\frac{3\epsilon}{1/4-\epsilon/2}\right)\geq\frac{31}{32}-\frac{\epsilon_1}{2}. \tag{4.14}$$

We henceforth assume that inequality 4.14 holds. This together with 4.13 yields:

$$\frac{N_{5<}(v)}{n_5}\geq\frac{31}{32}-\frac{\epsilon_1}{2}.$$

Hence E^n holds. We have just proved that if $\epsilon_1>0$ and $\epsilon>0$ are chosen so that 4.14 holds, then B_0^n, B_1^n, B_3^n and C^n together imply the event E^n.

■

Let $N_1(v)$ denote in the alignment v, the total number of subsequent pairs of ones in X satisfying all of the following conditions:
-The ones are both matched, that is aligned by v with a one from Y.
-Between the pair of ones in the X text there is exactly one zero and nothing else.
-The pair of ones in the Y text with which our pair from X is aligned, should contain only zeros in between them.
More precisely: for $v\in V^k$, we define:

$$N_1(v):=|\{i<k|v_i=(0,0),\pi_v(i+1)-\pi_v(i)=2\}|$$

where $v:=(v_1,v_2,\ldots,v_k)$.
Let $N_{1>}(v)$ denote in the alignment v, the total number of subsequent pairs of ones in X satisfying all of the following conditions:
-The ones are both matched, that is aligned by v with a one from Y.
-Between the pair of ones in the X text there are only zeros and we require that there is exactly one zeros.
-The pair of ones in the Y text with which our pair from X is aligned, should contain only zeros in between them and contained two or more.
More precisely:

$$N_{1>}(v):=|\{i<k|v_i=(0,0),\pi(i+1)-\pi(i)=2,\nu_v(i+1)-\nu_v(i)>2\}|.$$

Let B_4^n be the event that in the sequence X there are at least $((1/4)-\epsilon/16)n$ blocks of zeros of length two.

Let $p_1(v)$ be the conditional probability on X, that when we pick a block of one zeros at random in X, this block happens to satisfy the following two conditions:
1) The block is contained between two consecutive matched ones. (Matched by the alignment v.) This means that between the two matched ones there is our block of length one and nothing else.
2) The pair of ones in the Y-text to which the pair of consecutive ones are aligned by v contains only zeros in between them and at least two of them.
In other words, $p_1(v)$ designates the conditional probability (conditional on X) that when we pick at random a block of zeros of length one in X, there exists $i \leq k$, such that the randomly selected block is equal to $[\pi_v(i-1)+1, \pi_v(i)-1]$ and all of the following holds:

$$\pi_v(i) - \pi_v(i-1) = 2,$$

$$\eta_v(i) - \eta_v(i-1) > 2$$

and $|v_i| = 0$, where

$$v = (v_1, v_2, \ldots, v_k) \in (\mathbb{N} \times \mathbb{N})^k.$$

Recall that the total number of blocks of one zeros in X is denoted by n_1. Hence the conditional probability $p_1(v)$ is equal to:

$$p_1(v) := \frac{N_{1>}(v)}{n_1}.$$

Let us go back to our numerical example. Again, let $X = 101010101000001$ and let $Y = 101000110001$. Let v be the alignment

1	0	1	0			1	0	1	1	0	0	0	0	0	1
1	0	1	0	0	0	1			1	0	0	0			1

There are three blocks of zeros of length one in X. Hence, $n_1 = 3$.
Among the blocks of one zero in X only two are directly comprised between matched ones. Hence, $N_1(v) = 2$.
One of the two “suitable” blocks of one zero in X is matched with a block of zeros in Y strictly longer than 1. Hence, $N_{1>} = 1$ and $p_1(v) = 1/3$.

Let F^n designate the event that

$$p_1(v) \geq \frac{1}{4} - \frac{\epsilon_1}{2}.$$

for every optimal alignment $v \in V$.
Let D^n denote the event that for all $v \in V$ which is an optimal alignment, we have that

$$\frac{N_{1>}(v)}{N_1(v)} \geq \frac{1}{4} - \epsilon_1/4.$$

Lemma 4.4 *Assume that $\epsilon > 0$ and ϵ_1 satisfy inequality 4.20, then*

$$B_0^n \cap B_1^n \cap B_4^n \cap D^n \subset F^n \tag{4.15}$$

for all n.

Proof. We have for any alignment $v \in V$, that

$$\frac{N_{1>}(v)}{n_1} = \frac{N_{1>}(v)}{N_1(v)} \cdot \frac{N_1(v)}{n_1}. \tag{4.16}$$

When, the event D^n holds, then, for every $v \in V$, we have

$$\frac{N_{1>}(v)}{N_1(v)} \geq \frac{1}{4} - \epsilon_1/4.$$

Combining the last inequality with equality 4.16, yields

$$\frac{N_{1>}(v)}{n_1} \geq \left(\frac{1}{4} - \epsilon_1/4\right) \cdot \frac{N_1(v)}{n_1}. \tag{4.17}$$

We have that B_0^n and B_1^n together imply B_2^n. Hence, when B_0^n and B_1^n both hold, then every optimal alignment is in V. Hence, inequality 4.17 also holds for every optimal alignment v.
If B_0^n and B_1^n both hold, there are at most $(\epsilon n/16) + (\epsilon n/8)$ non-matched ones in each text X and Y for any optimal alignment v. This also implies that the number of blocks of on zeros in X which do not satisfy the criteria to be counted towards $N_1(v)$ is at most $2[(\epsilon n/16) + (\epsilon n/8)]$, (for every optimal alignment v). This implies that for every optimal alignment v, when B_0^n and B_1^n both hold, then

$$N_1(v) \geq n_1 - 3n\epsilon/8.$$

The last inequality with inequality 4.17 implies

$$\frac{N_{1<}(v)}{n_1} \geq \left(\frac{1}{4} - \epsilon_1/4\right) \cdot \left(1 - \frac{3n\epsilon}{8n_1}\right), \tag{4.18}$$

for every optimal alignment v. When B_4^n holds, we have that

$$n_1 \geq ((1/4) - \epsilon/16)n,$$

and hence

$$1 - \frac{3n\epsilon}{8n_1} \geq 1 - \frac{3\epsilon}{2 - \epsilon/2}.$$

Note that the expression on the right side of the last inequality goes to 1 as ϵ converges to zero. Using the last inequality in inequality 4.18 gives:

$$\frac{N_{1>}(v)}{n_1} \geq \left(\frac{1}{4} - \epsilon_1/4\right) \cdot \left(1 - \frac{3\epsilon}{2 - \epsilon/2}\right), \tag{4.19}$$

Note that

$$\lim_{\epsilon \to 0} \left(1 - \frac{3\epsilon}{2 - \epsilon/2}\right) = 1.$$

Hence for any $\epsilon_1 > 0$ fixed, if we chose $\epsilon > 0$ small enough, (how small depends on ϵ_1), we get

$$\left(\frac{1}{4} - \epsilon_1/4\right) \cdot \left(1 - \frac{3\epsilon}{1/4 - \epsilon/2}\right) \geq \frac{31}{32} - \frac{\epsilon_1}{2}. \tag{4.20}$$

We henceforth assume that inequality 4.20 holds. This together with 4.19 yields:

$$\frac{N_{1>}(v)}{n_1} \geq \frac{1}{4} - \frac{\epsilon_1}{2}.$$

Hence F^n holds. We have just proved that if $\epsilon_1 > 0$ and $\epsilon > 0$ are chosen so that 4.20 holds, then B_0^n, B_1^n, B_4^n and D^n together imply the event F^n.

■

Let us give a numerical example. Let $X = 101010101000001$ and let $Y = 101000110001$. Let v be the alignment

1	0	1	0			1	0	1	1	0	0	0	0	0	1
1	0	1	0	0	0	1			1	0	0	0			1

Note that there are two consecutive aligned ones with one zero in between in the X-part. Hence $N_1(v) = 2$. Note, that the third and fifth one in the string X are consecutive aligned ones with one zero between them. But there is also an non-aligned one between them, so they don't count towards $N_1(v)$. Instead the first consecutive aligned one counting towards $N_1(v)$ is given by the first and the second one in X. Then, the second and third one in X constitutes such consecutive couple of ones counting towards $N_1(v)$.
among the two pairs of consecutive aligned ones with one zero in between, there is one which has strictly more than 1 zero in between the ones in the Y part. Hence, $N_{1>} = 1$.
There is here one consecutive pair of aligned ones with five zeros in between. it is given by the fifth and sixth one in X. Hence, $N_5(v) = 1$.
The only pair of consecutive aligned ones with five zeros in between in the X part has 3 zeros in between in the Y part. Hence, $N_{5<} = 1$.

Recall that A^n is the event that X and Y are such that the inequalities 2.5 and 2.6 are satisfied.

The next combinatorial lemma of this paper is:

Lemma 4.5 *When $\epsilon_1 > 0$ satisfies inequality 4.24, then*

$$E^n \cap F^n \subset A^n \tag{4.21}$$

holds for every $n \in \mathbb{N}$.

Proof. We say that a block of zeros in X is *aligned with another block of zeros* in Y by v, if these blocks are in between consecutive mutually aligned ones.

Let us give an example. Let $x = 101$ and $Y = 1001$. Let v be the alignment

1	0		1
1	0	0	1

In this example X has one block of zeros. This block has length one. The text Y has also one block of zeros. This block has length two. We say that the block of zeros in X gets aligned by v with the block of zeros in Y.

Recall that to obtain $\tilde{L}_n$ from L_n, we pick a block of five zeros at random in X and reduce its length by one. Then we add a zero to a randomly chosen block of one zero in X. The modified text is denoted by $\tilde{X}$. The optimal score between $\tilde{X}$ and Y is $\tilde{L}$.
For any optimal alignment v, we have:
When the selected block of five zeros in X is aligned by v with a block of length strictly less than five, then the score is not reduced. When, on top of that, the extra zero is added to a block of length one which is aligned by v to a block of at least two zeros, than the score is increased by one. The block of length five and the block of length one are chosen independently from each other, so that we find:

$$P(\tilde{L}_n - L_n = 1|X,Y) \geq p_5(v) \cdot p_1(v) \tag{4.22}$$

for any optimal alignment v. When, E^n and F^n both hold, then, for any optimal alignment v we have:

$$p_5(v) \cdot p_1(v) \geq \left(\frac{1}{4} - \frac{\epsilon_1}{2}\right)\left(\frac{31}{32} - \frac{\epsilon_1}{2}\right) = \frac{1}{4} \cdot \frac{31}{32} - \epsilon_1 \left(\frac{35}{64} + \frac{\epsilon_1}{4}\right) \tag{4.23}$$

For $\epsilon_1 > 0$ small enough, we have that

$$\frac{35}{64} + \frac{\epsilon_1}{4} < 1. \tag{4.24}$$

We assume that ϵ_1 is chosen small enough so that 4.24 holds. Then, 4.22 and 4.23 together imply that

$$P(\tilde{L}_n - L_n = 1|X,Y) \geq \frac{1}{4} \cdot \frac{31}{32} - \epsilon_1 \tag{4.25}$$

Let v be an optimal alignment. We saw that when the selected block of length five is aligned with a block of length strictly less, than five then the score can not decrease, i.e.:

$$\tilde{L}_n - L_n \geq 0.$$

It follows that for any optimal alignment v, we have

$$P(\tilde{L}_n - L_n = -1|X,Y) \leq 1 - p_5(v) \tag{4.26}$$

When E_n holds, we have

$$p_5(v) \geq \frac{31}{32} - \frac{\epsilon_1}{2} \tag{4.27}$$

Together, the inequalities 4.26 and 4.27 imply

$$P(\tilde{L}_n - L_n = -1|X,Y) \leq \frac{1}{32} + \epsilon_1 \tag{4.28}$$

We just prove that when E^n and F^n both hold, then 4.25 and 4.28 both hold. In other words, we proved that E^n and F^n jointly imply A^n, when $\epsilon_1 > 0$ satisfies 4.24.
■

5 Probability bounds

5.1 The bounds

For an event B, we have that B^c denotes the complement of B.
We start with some useful bounds:

Lemma 5.1 *We have that there exists* $\gamma_0 > 0$ *not depending on* n *such that*

$$P(B_0^n) \geq 1 - e^{-\gamma_0 n}. \tag{5.1}$$

Note that γ_0 *depends on* ϵ.

Proof. The sequence X, resp. Y is i.i.d.. The probability that $X_i = 1$, resp. $Y_i = 1$ is equal to 0.5. Hence, by large deviation, the probability that the average $\sum_i^n X_i/n$ is different from its mean by more than $\epsilon/16$ is exponentially small in n. ■

The next bound is:

Lemma 5.2 *There exists* $\gamma_3 > 0$ *not depending on* n*, such that*

$$P(B_3^n) \geq 1 - e^{-\gamma_3 n}. \tag{5.2}$$

(Note that γ_3 *depends on* ϵ*).*

Proof. The blocks in X are i.i.d.. The probability that a block has length 5 is equal to 1/32. Large deviation applies. to the probability that the proportion of blocks of zeros which has length five is below expectation by $\epsilon/16$. Hence that probability is exponentially small in n. ■

Similarly to the last lemma:

Lemma 5.3 *There exists* $\gamma_4 > 0$ *not depending on* n*, such that*

$$P(B_4^n) \geq 1 - e^{-\gamma_4 n}. \tag{5.3}$$

(Note that γ_4 *depends on* ϵ*).*

Proof. Essentially the same as the proof of the previous lemma above. ■

The next lemma gives an upper bound on the number of elements in the set V.

Lemma 5.4 *We have that*

$$|V^k| \leq e^{H(\epsilon/4)k} 2^{\epsilon k/2}. \tag{5.4}$$

Proof. First we need to determine which entries are not zero. There are at most $\epsilon k/2$ non-zero entries, which have to be chosen from $2k$ entries. Hence this gives a total of

$$\binom{2k}{\epsilon k/2} \leq e^{H(\epsilon/4)} \tag{5.5}$$

possibilities. Then, we have to chose how large each entry is. This means that we have to distribute among the non-zero entries (which we already determined in the previous step) a total of $\epsilon k/2$ integer points. This is the same as finding an integer partition of the interval $2^{\epsilon k/2}$. There are at most $2^{\epsilon k/2}$ integer partitions of the interval $[0, \epsilon k/2]$. This with 5.5 means that V^k contains no more than

$$e^{H(\epsilon/4)} 2^{\epsilon k/2}$$

elements. ■

Eventually, we have

Lemma 5.5 *Assume that $\epsilon > 0$ and ϵ_1 are such that 5.8 and 5.13 both hold. Then, we have that:*

$$P(C^n) \geq 1 - e^{-\gamma_c n}, \tag{5.6}$$

where γ_c is a positive constant not depending on n, (but depending on ϵ).

Proof. Let C^n_* denote the event that if $N_5(v) \geq n/33$, then C^n holds. In other words

$$C^n_* = \left(C^n \cap \left\{N_5(v) \geq \frac{n}{33}\right\}\right) \cup \left\{N_5(v) < \frac{n}{33}\right\}.$$

Note that when B^n_3 holds there are at least $(1/32 - \epsilon/16)n$ blocks of five zeros in X. When B^n_0 and B^n_1 both hold, then we argued that there at most $6\epsilon n/16$ ones not matched total in both sequences X and Y for any optimal alignment v.
This gives that when B^n_0, B^n_1 and B^n_3 all hold, then for any optimal alignment v, we have that

$$N_5(v) \geq (1/32 - 7\epsilon/16)n. \tag{5.7}$$

For $\epsilon > 0$ small enough, we have

$$\frac{1}{32} - \frac{7\epsilon}{16} \geq \frac{1}{33}. \tag{5.8}$$

From here on we assume that 5.8 holds, so that 5.7 implies that

$$N_5(v) \geq \frac{n}{33}.$$

Hence

$$B^n_0 \cap B^n_1 \cap B^n_3 \subset \left\{N_5(v) \geq \frac{n}{33}\right\}$$

and hence

$$\left\{N_5(v) < \frac{5}{33}\right\} \subset B^{nc}_0 \cup B^{nc}_1 \cup B^{nc}_3 \tag{5.9}$$

We have that

$$C^{nc} \subset C^{nc}_* \cup \left\{N_5(v) < \frac{5}{33}\right\}$$

and hence with 5.9 and with lemma 4.1, we find

$$C^{nc} \subset C^{nc}_* \cup B^{nc}_0 \cup B^{nc}_3.$$

The last inclusion implies that

$$P(C^{nc}) \leq P(C_*^{nc}) + P(B_0^{nc}) + P(B_3^{nc}).$$

We already prove exponentially negative bounds for $P(B_0^{nc})$ and for $P(B_3^{nc})$. Hence it only remains to prove an exponential negative upper bound for $P(C_*^{nc})$.
Let $Z_1, Z_2, \ldots$ denote a sequence of iid geometric random variables with parameter $1/2$. Let W_i be the indicator variable which is equal to 1 if $Z_i < 5$. Hence, $P(W_i = 1) = 31/32$. We have that:

$$P(C_*^{nc}) \leq \sum_{k=n/33}^{\infty} |V^k| \cdot P\left(\frac{W_1 + W_2 + \ldots + W_k}{k} < \frac{31}{33} - \epsilon_1/4.\right) \tag{5.10}$$

By large deviation we find

$$P\left(\frac{W_1 + W_2 + \ldots + W_{n/33}}{n/33} < \frac{31}{33} - \epsilon_1/4.\right) \leq e^{-n\gamma(\epsilon_1)}, \tag{5.11}$$

for some constant $\gamma(\epsilon_1)$.
Combining 5.4 and 5.11, we obtain

$$|V^k| \cdot P\left(\frac{W_1 + W_2 + \ldots + W_k}{k} < \frac{31}{33} - \epsilon_1/4.\right) \leq e^{H(\epsilon)k} 2^{\epsilon k} \cdot e^{-k\gamma(\epsilon_1)}. \tag{5.12}$$

inequality 5.10 and 5.12 together provide an exponential upper bound for $P(C_*^{nc})$ as soon as the following inequality:

$$H(\epsilon) + \epsilon - \gamma(\epsilon_1) < 0 \tag{5.13}$$

is satisfied. Note that

$$\lim_{\epsilon \to 0} (H(\epsilon) + \epsilon) = 0,$$

whilst $\gamma(\epsilon_1) > 0$ for every $\epsilon_1 > 0$. This implies that for any $\epsilon_1 > 0$ fixed, we can have inequality 5.13 hold, by taking $\epsilon > 0$ small enough.

■ In complete similarly to the previous lemma we have:

Lemma 5.6 *Assume that $\epsilon > 0$ and ϵ_1 are such that 5.8 and 5.13 both hold. Then, we have that there exists $\gamma_d > 0$ not depending on n such that*

$$P(D^n) \geq 1 - e^{-\gamma_d n}. \tag{5.14}$$

Proof. This proof goes like the previous one and so is omitted. ■

5.2 Overview

In section 3 we prove that $\mathrm{Var}L_n$ is of order n. For this we assumed that Theorem 2.2 holds. So it still remains to prove theorem 2.2. First, let us mention that the order n for $\mathrm{Var}[L_n]$ follows from two things:

a) $\Delta_n := \tilde{L}_n - L_n$ needs to have a positive bias. More precisely, we want that for any $(x, y) \in A^n$ and any n, we have that

$$E[\tilde{L}_n - L_n | X = x, Y = y] > cst,$$

where $cst > 0$ is any positive constant.

b) The probability $P(A^n)$ needs to be close to one. More precisely we want the quantity $1 - P(A^n)$ to be no more than stretched negative exponential.

Let us first mention problem a). Note that between $\tilde{L}_n$ and L_n the score can change by at most one. In other words,

$$P(\tilde{L}_n - L_n \in \{-1, 0, 1\}) = 1. \tag{5.15}$$

From equation 5.15 and with the help of 2.5 and 2.5, we find:

$$E[\tilde{L}_n - L_n = 1 | X = x, Y = y) \geq \frac{31}{32} - \frac{1}{4} - \epsilon_1 = \frac{23}{32} - \epsilon_1, \tag{5.16}$$

Taking $\epsilon_1 > 0$ small enough so that

$$\epsilon_1 < \frac{23}{32} \tag{5.17}$$

Insures the positive bias.

Let us next discuss problem b). In the last section, we proved that the following inclusions hold:
$B_0^n \subset B_1^n$, when condition 4.3 holds.
$B_0^n \cap B_1^n \subset B_2^n$, when condition 4.8 holds.
$B_0^n \cap B_1^n \cap B_3^n \cap C^n \subset E^n$, when condition 4.14 holds.
$B_0^n \cap B_1^n \cap B_4^n \cap D^n \subset F^n$, when condition 4.20 holds.
$E^n \cap F^n \subset A^n$, when 4.24 holds.
These inclusions imply, when the conditions 4.3, 4.8,4.14, 4.20 and 4.24 all hold, that

$$B_0^n \cap B_3^n \cap B_4^n \cap E^n \cap D^n \subset A^n,$$

and hence

$$P(A^{nc}) \leq P(B_0^{nc}) + P(B_3^{nc}) + P(B_4^{nc}) + P(C^{nc}) + P(D^{nc}) \tag{5.18}$$

Inequality 5.18 implies that $P(A^{nc})$ is negatively exponentially small in n, as soon as we have negative exponential bounds for $P(B_0^{nc})$, $P(B_3^{nc})$, $P(B_4^{nc})$, $P(C^{nc})$ and $P(D^{nc})$.
The probabilities $P(B_0^{nc})$, $P(B_3^{nc})$ and $P(B_4^{nc})$, only depend on ϵ. For any value of $\epsilon > 0$, the inequalities 5.1, 5.2 and 5.3 provide negative exponential bounds for $P(B_0^{nc})$, $P(B_3^{nc})$ and $P(B_4^{nc})$. Hence, there is no special condition on $\epsilon > 0$ and $\epsilon_1 > 0$ in order to ensure that $P(B_0^{nc})$, $P(B_3^{nc})$ and $P(B_4^{nc})$ are negative exponentially small in n.
We also proved exponential negative upper bounds for $P(C^{nc})$ and $P(D^{nc})$. These

bounds are the inequalities 5.6 and 5.14. However these bounds only hold if ϵ and ϵ_1 satisfy all the conditions 5.8, 5.13, 5.8 and 5.13.
To prove that $P(A^{nc})$ is exponentially small it thus remains to prove that there exists ϵ and ϵ_1 satisfying all the following:
a) The condition 5.17, which gives $\tilde{L}_n - L_n$ its conditional bias 5.16.
b) All the conditions for the inclusions. These is are the inequalities 4.3, 4.8, 4.14, 4.20 and 4.24.
c) The conditions for the exponential upper bounds for $P(C^{nc})$ and $P(D^{nc})$. These are the inequalities 5.8, 5.13, 5.8 and 5.13.
To see that there exists $\epsilon, \epsilon_1 > 0$ satisfying all the above conditions simultaneously, note that these conditions can be classified in three types:
type I: conditions involving only ϵ_1. These conditions all hold for $\epsilon_1 > 0$ small enough.
Type II: conditions involving ϵ_1 and ϵ. All these conditions are such that for any $\epsilon_1 > 0$ fixed, they hold as soon as $\epsilon > 0$ is small enough.
Type III: conditions involving only ϵ. They all hold as soon as ϵ is taken small enough.
It is now easy to see that there exists $\epsilon_1 > 0$ and $\epsilon > 0$ such that all the conditions 5.17, 4.3, 4.8, 4.14, 4.20, 4.24, 5.8, 5.13, 5.8 and 5.13 simultaneously hold. For this chose first and $\epsilon_1 > 0$ small enough so that all equations of type I are satisfied. Then chose $\epsilon > 0$ small enough so that all conditions of type II and type III are satisfied. Summarizing we have just shown that: There exist $\epsilon, \epsilon_1 > 0$ and $c_1, s > 0$ not depending on n, such that if $s(1,1) \geq s$, then

$$P(A^{nc}) \leq e^{c_1 \cdot n}, \tag{5.19}$$

for all n.

References

[1] D. Aldous and P. Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

[2] K. S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, 4(4):1074–1082, 1994.

[3] R. Arratia, L. Gordon, and M.S. Waterman. The Erdős-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.*, 18(2):539–570, 1990.

[4] R. Arratia and M.S. Waterman. The Erdős-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.*, 17(3):1152–1169, 1989.

[5] R. Arratia, L. Gordon, and M. Waterman. An extreme value theory for sequence matching. *Ann. Statist.*, 14(3):971–993, 1986.

[6] R. Arratia and M. S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.

[7] R.A. Baeza-Yates, R. Gavaldà, G. NaVarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.

[8] J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[9] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1999.

[10] V. Chvàtal and D. Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[11] P. Clote and R. Backofen. *Computational molecular biology.* Wiley Series in Mathematical and Computational Biology. John Wiley & Sons Ltd., Chichester, 2000. An introduction.

[12] V. Dančík and M. Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[13] J. G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[14] Hauser, Matzinger, and Martinez. Large deviation montecarlo method for lcs. submitted.

[15] M. Kiwi, M. Loebl, and J. Matousek. Expected length of the longset common subsequence for large alphabets. *preprint*, 2003.

[16] Lember and Matzinger. Deviation from mean in sequence comparison with a periodic sequence. submitted.

[17] C. Neuhauser. A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.*, 22(3):1603–1629, 1994.

[18] M. Paterson and V. Dančík. Longest common subsequences. In *Mathematical foundations of computer science 1994 (Košice, 1994)*, volume 841 of *Lecture Notes in Comput. Sci.*, pages 127–142. Springer, Berlin, 1994.

[19] P. A. Pevzner. *Computational molecular biology: An algorithmic approach* A Bradford Book. MIT Press, Cambridge, MA, 2000.

[20] S. Robin, F. Rodolphe, and S. Schbath. *ADN, mots et modèles*, Belin, Paris, 2003.

[21] J.M. Steele. Long common subsequences and the proximity of two random strings *SIAM J. Appl. Math.*, 42(4):731–737, 1982.

[22] J.M. Steele.An Efron-Stein inequality for nonsymmetric statistics *Ann. Statist.*, 14(2):753–758, 1986.

[23] M.S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.